PKU-NEC @ TRECVid 2012 SED: Uneven-Sequence Based Event Detection in Surveillance Video*

Ziwei Xia [a], ***Xiaoyu Fang*** [a],
Yaowei Wang [a], ***Wei Zeng*** [b], ***Hongming Zhang*** [b], ***Yonghong Tian*** [a+]
[a] National Engineering Laboratory for Video Technology, School of EE & CS, Peking University
[b] NEC Laboratories, China
[+] Corresponding author: Phn: +86-10-62758116, E-mail: yhtian@pku.edu.cn

Abstract

In this paper, we describe our system for interactive and retrospective surveillance event detection task in TRECVid 2012. We focus on pair-wise events (e.g., PeopleMeet, PeopleSplitUp, Embrace) that need to explore the relationship between two active persons, and action-like events (e.g. ObjectPut, CellToEar, PersonRuns and Pointing) that need to find the happenings of a person's action. Our team had participated in the TRECVid SED task from 2009 to 2011. This year the new improvements of our system are three-fold. First, we introduce geometric constraints into the human detection. Second, we propose a robust tracking-by-detection approach with an optimized observation mode to address ID switching and tracking drifting. Third, an uneven-sequence classifier is employed for action-like event detection. Overall, we have submitted three versions of interactive results, which are obtained by using different human detection, tracking and events detection modules, and one version of retrospective result. For the definition of interactive task is ambiguous, we mainly introduce our work on retrospective task in this paper. According to the results in the TRECVid SED formal evaluation, our experimental results of retrospective task are promising.

1. Introduction

Event detection in surveillance enviroments is a critical application in computer vision field. Although event detection in surveillance video has been much studied, the applicable system is still far away from our life due to the following challenges:(1)crowded scenes(2)various illuminations(3)heavy occlusion(4)low resolution(5)various human activities(6)unclear event definition(7)real-time computation(8)uneven distribution of events. In order to address part of the challenges mentioned above, our team, PKU_NEC, participated in the interactive and retrospective surveillance event detection task in TRECVid 2012.

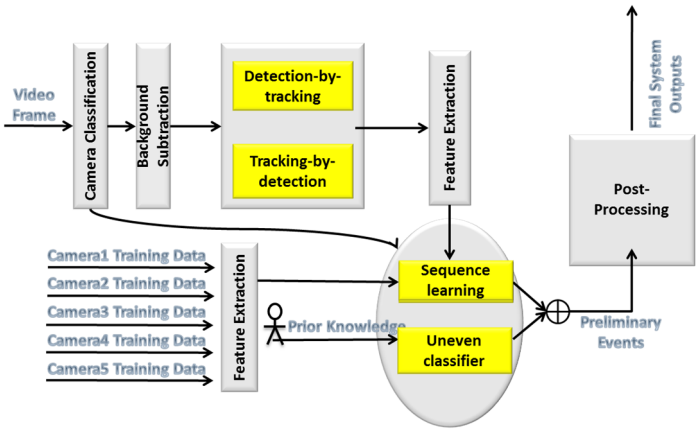

Fig.1 Diagram of our system

This year we classify seven events into two classes. One class is pair-wise events (e.g., PeopleMeet, PeopleSplitUp, Embrace) that need to explore the relationship between two active persons, the other is

* This work was cooperatively done by Peking University and NEC Laboratories, China. This work is partially supported by grants from the Chinese National Natural Science Foundation under contract No. 61035001, and No. 61072095, National Basic Research Program of China under contract No. 2009CB320906, and Fok Ying Dong Education Foundation under contract No. 122008.

action-like events (e.g. ObjectPut, CellToEar, PersonRuns and Pointing) that need to find the happening of a person's action. The diagram of our system is shown in Fig.1.

Three key improvements are made in the system than the 2011 and 2010 systems.

First, we introduce geometric constraints into the human detection. Geometric constraint is a widely used kind of contextual information which could be utilized to generate regions of interest (ROIs)[12][13]. With the ROIs, the search area of the detector could be limited to regions where human may appear. In this way, many background noises are dismissed. Experimental results show that our system can achieve a much better precision and recall than our previous systems.

Second, the ID switching or tracking drifting is the primary challenge of tracking task. To address these problems, we propose a robust tracking-by-detection approach with an optimized observation model. Our observation model of particle filter fuses the detection results and three states of trackers in a unified probabilistic framework, where each state is represented with a MIL classification model [16]. These states include the Original State that denotes the initial state of a tracker in the video, the Current State that characterizes the tracker's state at the present time, and the Max-Difference State that represents the state of the tracker which can best capture the appearance of the object and thus is most different from other trackers. Therefore, three state classifiers can be utilized to recognize the current status of each tracker. In our experiments, ID switching and tracking drift can be effectively avoided.

Third, an uneven-sequence classifier is employed for action-like event detection. We define some states and learn the transition relation among these states to detect the event. Then the data is processed by an uneven-sequence classifier. Experimental results show our system is feasible and effective. According to the results in the TRECVid SED formal evaluation, our experimental results are promising.

The remainder of this paper is organized as follows. In section 2, we describe our head-shoulder detection and tracking approach. In section 3, we present our approach for detecting different events and the different methods for processing interactive task and retrospective task. Experimental results and analysis are given out in section 4. Finally, we conclude this paper in section 5.

2. Detection and Tracking

2.1 Detection-by-Tracking and Tacking-by-Detection

Human detection is an important step in this system. For there are many occlusions in the TRECVid corpus, we simultaneously apply head-shoulder detection and human body detection in our approach. Many people in complex scenes will be occluded for a fairly long period. Thus, the human detection in individual frames and data-association of the detection results among several continuous frames are challenging and ambiguous. In [1] and [2], temporal coherency is involved to detection. In our system, we try to exploit temporal coherency by integrate detection and tracking in one unified framework. People-trajectories are extracted from a small number of consecutive frames and from those trajectories build models of the individual people.

Human Body Detection with geometric constraints

Traditional detection methods suffer the performance degradation caused by clutter background and are often very time-consuming. To solve these problems, we employ geometric constraints to generate regions of interest (ROIs). Then the search area of the detector could be limited to the regions where human may appears. In this way, we obtain the human body detection result much faster and more accurate than traditional detection methods.

Head-Shoulder Detection

In [3], Dalal and Triggs proved that Histograms of Oriented Gradients are powerful for human detection. In order to speed up, Zhu et al. [4] combined the cascaded rejection approach with HOG feature. They used AdaBoost to select the best features and constructed the rejection-based cascade.

In our system, we apply a simple and fast method to generate initial detection result. We use HOG feature to represent head-shoulder samples, and apply linear SVM classifier. With the coarse foreground regions extracted from background modeling module, we wipe out candidate regions that do not have enough foreground in them. Moreover, by using statistical data of each camera, we can simply estimate the possible size of person appeared in different positions. Thus, the detection process is more efficient.

In practice, we labeled about 5000 head-shoulders as positive training samples, and collected hundreds of images without head-shoulders as the source to extract negative training samples.

Head-Shoulder Detection Update

The final probability of detection $\mathrm{p}(d_N)$ of current frame N will be predicted or updated with the following equation

$$\mathrm{p}(d_N) = w_1 C(d_N) + w_2 S_f(d_N, d_{N-1}) + w_3 S_l(d_N, d_{N-1}),$$

where w_1, w_2, and $w_3$are weights, d_N is the detection in frame N, $C(d_N)$ is confidence of d_N, $S_f(d_N, d_{N-1})$ is the appearance similarity (HOG) of d_N and d_{N-1}, and $S_l(d_N, d_{N-1})$ is the location and scale similarity of d_N and d_{N-1}. $S_l(d_N, d_{N-1})$ is defined by

$$S_l(d_N, d_{N-1}) = p_N\left(\frac{size_N - size_{N-1}}{size_N}\right) \times p_N(|d_N - d_{N-1}|),$$

where $size_N$ is the size of d_N, $p_N\left(\frac{size_N - size_{N-1}}{size_N}\right)$ is the scale similarity of d_N and d_{N-1}, and $p_N(|d_N - d_{N-1}|)$ is the location distance of d_N and d_{N-1}.

We set different weights for different scenes. Head-shoulder detection updating will terminate when the tracking result change. Then, if the detection results have maximum $\mathrm{p}(d_N)$ and $\mathrm{p}(d_N) > Th$ (Th is the detection threshold) they are appended to the final detection results.

Considering most of the head-shoulders of human are small and blurred, we apply the online Multiple Instance Learning algorithm [6] instead of the Online Boosting algorithm in [5]. For each classifier, weak learners are selected using MIL Boost.

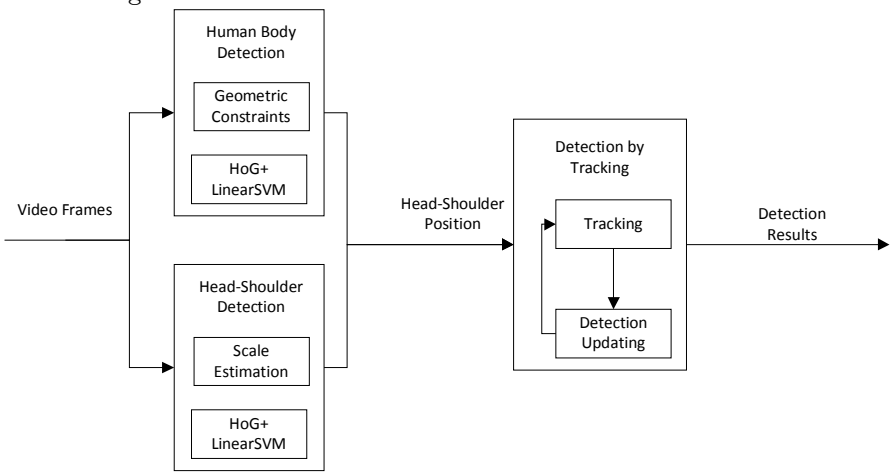

Fig.2 Framework of Detection-by-Tracking

Tracking by Detection with Optimized Observation Model

Our tracking algorithm is based on estimating the distribution of each target state by a particle filter. We use a constant velocity motion model of each particle [6]. To compute the weight for a particle of the tracker, we estimate the likelihood of each particle. For this purpose, we combine information from different sources, the associated detection score, the preliminary detection results of the detection-by-tracking algorithm mentioned in section 2.1, and the classifier outputs.

In the TRECVid corpus, target appearance always changes significantly. Based on the data association algorithm described by Michael D. Breitenstein[6], we relate no more than one detection result with a tracker. According to the detection result, we build a matching function $S(tr, d)$ for each pair (tr, d) of detection d and tracker tr. Then we associate its position and its size to get a set of function value. Finally according to Hungarian algorithm [18], we find out the most relevant detecting result of each tracker.

For each tracker we built a particle filter as [19] to predict the state of a tracker in next frame. The state includes the position and speed of a tracker. In order to track the states change of one tracker from beginning to the end, we preserve its three states, namely the Original state, Current state and Max-difference state. Then we propose a new model to fuse the three states of a tracker and the associated detection result.

2.2 Head-Shoulder Detection Based on Gradient Tree Boosting

We also propose another approach using Gradient Tree Boosting [7] to detect object with high accuracy and fast speed. The essential component of the proposed approach is a cascade Gradient Boosting Tree

based object detector, which uses HoG features as object representation. In order to track multiple objects in Trecvid video, we adopt Multiple Hypothesis Tracking (MHT) Method. MHT algorithm was invented by Reid [8] in the context of multi-target tracking, and was improved by Cox and Hingorani[9] by an efficient implementation.

We also propose another approach using Gradient Tree Boosting [8] to detect object with high accuracy and fast speed and adopting Multiple Hypothesis Tracking (MHT) Method.

Head-Shoulder Detection Based on Gradient Tree Boosting

Fig.3 shows the overall architecture of our object detection approach, which contains training stage and detection stage. The essential component of the proposed approach is a cascade Gradient Boosting Tree based object detector, which uses HoG (Histograms of Oriented Gradients) [4] features as object representation. During training stage, a lot of samples of object and negative images are used to select informative features and to train the object detector. The detection stage is the process to locate object instances in any given input image by using the object detector.

Gradient boosting method was invented by Jerome H. Friedman [8] in 1999 and can be used for classification problems by reducing them to regression with a suitable loss function. In our system, we use decision tree as base learner, and cascade gradient boosting as learning framework.

Multiple Hypothesis Tracking Method

In order to track multiple objects in Trecvid video, we adopt Multiple Hypothesis Tracking (MHT) Method. MHT algorithm was invented by Reid [9] in the context of multi-target tracking, and was improved by Cox and Hingorani [10] by an efficient implementation. It uses statistical data association to deal with some tracking issues, such as track initiation, track termination, and track continuation. In our system, head-shoulder detection is incorporated with MHT tracking process to construct one integrated system. For any video, the track results are computed frame by frame. We tested the system on Trecvid dataset. Table 2 shows the evaluation results.

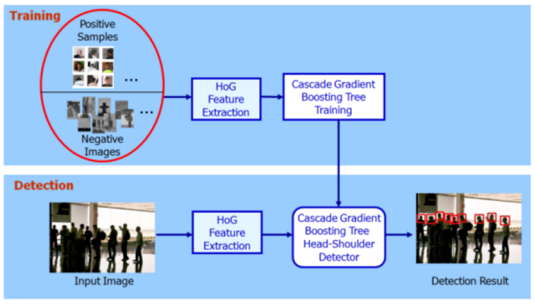

Fig.3 Object detection architecture based on Gradient Tree Boosting

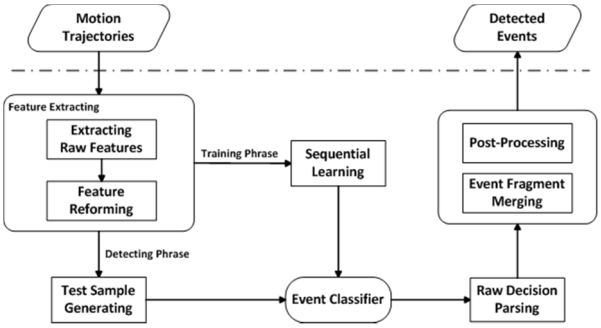

Fig.4 Flowchart of sequential learning based event detection

3. Event detection

In Trecvid 2012, surveillance event detection task has introduced a new task from last year, which is interactive surveillance event detection. The main difference of our approach for handling interactive SED task is in the discrimination state. The difference between interactive task and retrospective task can be described as follow: in the discrimination state, we spend no more than 25 minutes on one event. And the results, whose confidence is beyond the threshold value, would be considered final results to output.

3.1 Pair-wise Event Detection

To detect the pair-wise events in this year's SED task, the pair-wise events, such as PeopleMeet, PeopleSplitUp, and Embrace, are considered as a time-variant holistic pattern, and spatio-temporal cubic feature and sequence discriminant learning method are introduced to serve the detection task.

The discriminative patterns for these three events in video sequences are inherently time sequential. However, most pervious activity recognition methods did not handle this properly with only modeling the patterns in single frames or simply concatenating them together. In our solution, the event is considered as a whole sequence and described by the spatio-temporal cubic feature. Specifically, we employ Support Vector Machine with dynamic time alignment kernel proposed in [11]. This method handles time series

feature with varying length and the learning procedure is based on a maximum margin criterion. With the sequence discriminant learning method, the temporal correlations between different stages of the event are properly considered, and decisions based on integrated event sequences are reliable and semantically reasonable.

As shown in Fig.4, features are extracted based on the motion trajectories generated by human detecting and tracking module mentioned in previous sections. We use statistical trajectory descriptor to describe the relationship of the trajectories. Let $A_m = \left[a_1^1, ..., a_i^{\lceil i/L \rceil}, ..., a_T^{\lceil T/L \rceil}\right]$ and $B_n = \left[b_1^1, ..., b_i^{\lceil i/L \rceil}, ..., b_T^{\lceil T/L \rceil}\right]$ be motion trajectories of objects m and n, where a_i and b_i are tuples (x, y) of the object coordinates in 2D image plane at time i, and m, n are objects' identifiers. To represent the relationships of the objects in each cube and remove the influences of occasional error caused by detection and tracking, statistical data is employed, such as mean distance one from another, mean relative speed magnitude, and mean overlapped area of objects' regions. Meanwhile, the difference of these statistical data between current and previous cubes is important as well. Therefore, trajectory descriptor of k^{th} cube is extracted as follows:

$$TD^k = \left\{c_{dis}^k, c_{sp}^k, c_{ov}^k, dc_{dis}^k, dc_{sp}^k, dc_{ov}^k\right\}$$

where c_{dis}^k, c_{sp}^k and c_{ov}^k is mean distance, mean relative speed magnitude and mean overlapped area within k^{th} cube respectively, and

$$dc_{dis}^k = \begin{cases} 0, & k = 1 \\ c_{dis}^k - c_{dis}^{k-1}, & k > 1 \end{cases}$$

$$dc_{sp}^k = \begin{cases} 0, & k = 1 \\ c_{sp}^k - c_{sp}^{k-1}, & k > 1 \end{cases}$$

$$dc_{ov}^k = \begin{cases} 0, & k = 1 \\ c_{ov}^k - c_{ov}^{k-1}, & k > 1 \end{cases}$$

The total process is described as follow: We first segment video sequences into several cubes, and then, according to the locations of every person in a frame, we calculate the mean absolute velocity, acceleration, distance between each pair of people and the angular separation of moving directions in each cube as the raw features. Then the extracted raw features from the same video clips (ground truth event samples for training and test samples for detecting) are transformed to structural sequence feature. Some statistics of raw features are also included into the reformed features to explicitly employ the information of the temporal dependencies over adjacent frames.

With the structural features, an appropriate implementation of SVM with dynamic time alignment kernel [8], is applied to train events classifiers and make decisions. As the raw decision is a sequence of binary decisions for each frame in a testing sample, we need to parse it into a single decision for the testing sample with the strategy like voting. As the detection task is actually transformed to a classification problem by using sliding window method to generate testing samples, the original results would be fragmental. So in the post-processing phrase, we merge the preliminary detections and introduce some prior knowledge based rules to filter out incredible detections. These rules are usually empirical restrictions such as a distance threshold between persons before "PeopleSplitUp" or after "PeopleMeet".

3.2 Action-like Event Detection

To detect action-like events, such as "ObjectPut", "CellToEar", "PersonRuns" and "Pointing", an uneven-sequence classifier is employed for action-like event detection. We first define some states and learn the transition relation among these states. Then a state transition model is constructed for each event. Base on the tracking results of objects, we use histogram of optical flow (HOF) for "ObjectPut" and MoSift[21] for other events to represent their motions, which will cause transition of their states. Then we utilize an uneven-sequence classifier to discriminate the events. The reason we don't apply original SVM is that the performance of original SVM is quite poor when to use for event detection in this surveillance video. That is because there are two challenges in the event detection.

First, the events of greatest semantic interest for detection are often rare compared to the typical events. That is to say, the distribution of the event data is highly uneven. As showed in Fig. 5, Pointing event (red) also co-occur with numerous typical event (yellow). There are a lot of unconcerned instances with only a few interest event instances in Trecvid dataset, and using normal classifiers (e.g., original SVM) to detect rare

events usually results in low detection rate.

Fig.5 Pointing event in TRECVid.

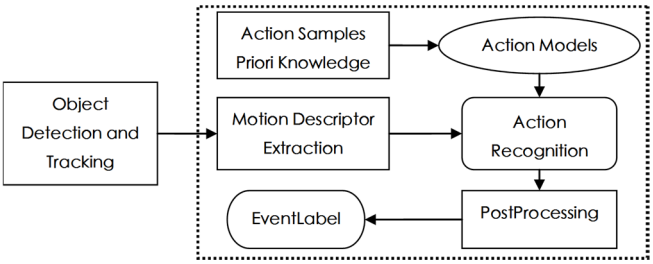

Fig.6 Action-like events detection

Second, the events are different in the durations of time because of the variety of the events. While SVM assumes that each sample is a vector of fixed dimension, and hence it can't deal with the variable length sequences directly. For this reason, most of the efforts that have been made so far to apply SVM to event detection employ linear time normalization, where input feature vector sequences with different lengths are aligned to same length. A variant of this approach is to employ HMM (hidden Markov model) into the whole event detection system, such as S-HSMM [23], Fisher kernels [24] and conditional symmetric independence (CSI) kernels [25]. Another approach is to incorporate the operation of dynamic time alignment into the kernel function itself, like DTAK-SVM [22]. Since HMMs can treat sequential patterns, SVM that employs the models based on HMMs can handle sequential patterns as well.

To address these challenges, we employ Uneven-Sequence SVM into the event detection system.

Sequence Learning

Suppose there are two vector sequences X and V, which may have different sequence length. Hence, we introduce a new class of kernel, named Dynamic Time Alignment Kernel [22], to find the optimal path that maximizes the accumulated similarity. It defined as follow:

$$K_s(X,V) = \max_{\psi,\theta} \frac{1}{M_{\psi\theta}} \sum_{i=1}^{N} m(i) K(x_{\psi(i)}, v_{\theta(i)})$$

subject to

$$1 \le \psi(i) \le \psi(i+1) \le |X|, \psi(i+1) - \psi(i) \le Q$$
$$1 \le \theta(i) \le \theta(i+1) \le |V|, \theta(i+1) - \theta(i) \le Q$$

where $m(i)$ is a nonnegative path weighting coefficient, N is the length of the warping path, ψ and θ stand for a warping path, and Q is a constant constraining the local continuity. In this paper, we use Radial Basis Function (RBF) kernel as K.

Processing Uneven Data

Although we have introduced the dynamic time-alignment kernel into the SVM model, it has been noticed that the performances of classifier is still quite poor because of the highly uneven dataset.

In [20], the authors proposed the uneven margins of SVM model, named SVM with Uneven Margins, to cope with binary classification problems where classes are highly unbalanced. When introducing a margin parameter into the optimization problem of the SVM to set the positive margin be some larger than the negative margin in the SVM, we can obtain the following optimization problem (OP2):.

$$\min imise_{w,b,\xi} \langle w, w \rangle + C \sum_{i=1}^{l} \xi_i$$

subject to

$$\langle w, x_i \rangle + \xi_i + b \ge 1 \quad if\ y_i = +1$$
$$\langle w, x_i \rangle - \xi_i + b \le -\tau \quad if\ y_i = -1$$
$$\xi_i \ge 0 \quad for\ i = 1, \ldots, m$$

4. Experiment and results

Our team submitted three versions of results, which are obtained by using different human detection, tracking and events detection modules, and one version of retrospective result.

Table 1 Detection results of this year and last year

Camera1	**Recall**	**Precision**	**F-score**	**Camera2**	**Recall**	**Precision**	**F-score**
Last Year	0.557	0.848	0.6724	Last Year	0.372	0.785	0.5048
This Year	0.587	0.832	0.6883	This Year	0.406	0.792	0.5368
Camera3				**Camera5**			
Last Year	0.423	0.756	0.5425	Last Year	0.318	0.775	0.4510
This Year	0.402	0.784	0.5314	This Year	0.388	0.762	0.5142

Table 2 Tracking results of this year and last year

Camera1	**MOTA**	**MOTP**	**Miss**	**FA**	**ID Switch**
Last Year-MHT	0.368	0.571	0.486	0.134	0.012
Last Year-PFT	0.364	0.567	0.472	0.154	0.010
This Year	0.396	0.562	0.378	0.115	0.012
Camera2					
Last Year-MHT	0.151	0.601	0.680	0.160	0.009
Last Year -PFT	0.213	0.607	0.644	0.132	0.011
This Year	0.261	0.634	0.535	0.198	0.007
Camera3					
Last Year-MHT	0.198	0.583	0.680	0.160	0.009
Last Year-PFT	0.271	0.591	0.667	0.050	0.010
This Year	0.290	0.617	0.624	0.075	0.011
Camera5					
Last Year-MHT	0.168	0.591	0.737	0.088	0.008
Last Year-PFT	0.170	0.589	0.731	0.089	0.009
This Year	0.181	0.632	0.702	0.107	0.010

Table 1 and 2 show the comparison detection and tracking results between the best outputs of our system this year and those of last year in TRECVid SED. It can be seen from the tables that detection result is improved greatly in recall with low or no decrease in the precision. Here we introduce Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) [8], metrics used in PETS 2009, to evaluate overall performance. These ID switches used in MOTA are calculated from the number of identity mismatches in a frame, from the mapped objects in its preceding frame. The MOTP is calculated from the spatiotemporal overlap between the ground truth tracks and the algorithm's output tracks. Conclusion can be drawn from table 2 that our performance is improved greatly. Moreover, due to the use of ROIs based on geometric constraints, our average detection speed is increased greatly from 0.6 fps to 10 fps.

For the part of event detection, we firstly show the performance of our approach on TRECVid 2007 dataset. The comparison results of action-like and pair-wise event detection on TRECVid 2007 dataset are respectively shown in Table 3 and Table 4.

Table 3 Comparison results of action-like event detection on TRECVid 2007 dataset

CellToEar	**#Ref**	**#Sys**	**#CorDet**	**#FA**	**#Miss**	**Act.DCR**
SVM	77	125	1	124	76	1.0756
WS-JTM[26]	77	26	2	24	75	0.9912
Ours	77	43	5	38	72	0.9622
Pointing						
SVM	401	492	16	476	385	1.3001
WS-JTM[26]	401	317	34	283	367	1.1174
Ours	401	107	35	72	366	0.9641

Table 4 Results of pair-wise event detection using different features on TRECVid 2007 dataset

PeopleMeet	**#Ref**	**#Sys**	**#CorDet**	**#FA**	**#Miss**	**Act.DCR**
BoW[27]	288	81	6	75	282	1.0327
Tr[28]	288	570	7	16	281	0.9871
Concatenation[17]	288	547	7	15	281	0.9864
Ours	288	453	6	2	282	0.9806
Embrace						
BoW[27]	74	73	1	15	73	0.9972
Tr[28]	74	1085	2	2	72	0.9744
Concatenation[17]	74	141	4	42	70	0.9759
Ours	74	523	7	54	67	0.9440

PeropleSplitUp						
BoW[27]	171	56	1	4	170	0.9970
Tr[28]	171	491	23	141	148	0.9662
Concatenation[17]	171	334	23	135	148	0.9619
Ours	171	234	34	137	137	0.8990

In Table 3, we compare the results with two approaches, original SVM and WS-JTM [26]. From this table, we can see that the performance of original SVM is much higher than other approaches. That is to say original is not suitable for event detection in this dataset. And our approach outperforms the two approaches, especially for the Pointing event. From the results in Table 3, US-SVM is demonstrated to be successful for event detection in surveillance video. Compared to the WS-JTM approach, our approach will bring in less false positive on the basis of high classification accuracy. This is not surprising because the principle of our approach is to move the positive margin to detect more positive instances.

In Table 4, we compare the results of pair-wise event detection using different features. Our approach shows best performance in both experiments. On contrast, the concatenated form feature [17] may get worse performance than single type of descriptors, for its improper combination. It is proved that our method could find relatively appropriate parameters for feature fusion through optimizing classification performance.

Table 5 Comparison results of retrospective task between eSur and best outputs in 2012

PeopleMeet	**#Ref**	**#Sys**	**#CorDet**	**#FA**	**#Miss**	**Act.DCR**
2012's Best	449	2382	18	61	431	0.9799
2012's eSur	449	2382	18	61	431	0.9799
PeopleSplitUp						
2012's Best	187	976	84	892	103	0.8433
2012's eSur	187	167	8	64	179	0.9782
Embrace						
2012's Best	175	643	68	575	107	0.8
2012's eSur	175	5234	15	111	160	0.9507
ObjectPut						
2012's Best	621	50	8	34	613	0.9983
2012's eSur	621	50	8	34	613	0.9983
CellToEar						
2012's Best	194	491	15	248	179	1.004
2012's eSur	194	491	15	248	179	1.004
PersonRuns						
2012's Best	107	587	37	550	70	0.8346
2012's eSur	107	785	9	181	98	0.9752
Pointing						
2012's Best	1063	76	31	32	1032	0.9813
2012's eSur	1063	39120	32	74	1031	0.9942

Table 5 and Table 6 respectively show the comparison results of retrospective task and interactive task between eSur and best outputs this year.

According to the results in Table 5, our experimental results of retrospective are promising this year, especially for the events PeopleMeet, ObjectPut and CellToEar. It can be seen that PeopleMeet, ObjectPut and CellToEar have all outperformed other participators. The DCRs of the events, PeopleSplitUp, Embrace and PersonRuns, are higher than the best results, but the false alarms of them are much less. For the event Poingting, the correctly detected number is more than that of best results, and DCR of our Pointing is also comparable with the best of this year.

Table 6 Comparison results of interactive task between eSur and best outputs in 2012

PeopleMeet	**#Ref**	**#Sys**	**#CorDet**	**#FA**	**#Miss**	**Act.DCR**
2012's Best	449	230	77	153	372	0.8787
2012's eSur	449	143	22	121	427	0.9907
PeopleSplitUp						
2012's Best	187	173	46	127	141	0.7957
2012's eSur	187	143	8	39	179	0.97
Embrace						
2012's Best	175	181	72	109	103	0.6243
2012's eSur	175	164	13	151	162	0.9752
ObjectPut						
2012's Best	621	403	117	286	504	0.9054
2012's eSur	621	50	8	34	613	0.9983
CellToEar						
2012's Best	194	245	15	230	179	0.9981

2012's eSur	194	245	15	230	179	0.9981
PersonRuns						
2012's Best	107	114	48	66	59	0.573
2012's eSur	107	190	9	181	98	0.9752
Pointing						
2012's Best	1063	683	263	420	800	0.8903
2012's eSur	1063	91	29	62	1034	0.9931

The comparison results of interactive task between eSur and best outputs this year are shown in Table 6. It can be seen that the event CellToEar of our approach has outperformed other approaches. For the rest events, the approach, who attain the best results, has manually selected the correct detections as final result. And our approach automatically obtains the final result in a limited time (no more than 25 minutes). In our opinion, this is the main reason why our approach is beaten by the approach of best results.

Table 7 shows the comparison results of retrospective task between eSur this year and last three years. In this table, "non-participation" means that we didn't participate in the TRECVid SED that year. We have participated in the events PeopleMeet, PeopleSplitUp and Embrace for four years. And the performances of PeopleMeet and Embrace have been almost improved year by year. As shown in this table, the performances of the rest events are improved to a certain degree compared to the previous system. According to the results of Table 7, our approach is verified to be feasible.

Table 7 Comparison results of retrospective task between eSur this year and last three years

PeopleMeet	**#Ref**	**#Sys**	**#CorDet**	**#FA**	**#Miss**	**Act.DCR**
2009's eSur	449	125	7	118	442	1.023
2010's eSur	449	156	12	144	437	1.02
2011's eSur	449	2382	24	108	125	0.9820
2012's eSur	449	2382	18	61	431	0.9799
PeopleSplitUp						
2009's eSur	187	198	7	191	180	1.025
2010's eSur	187	167	16	136	171	0.959
2011's eSur	187	2988	4	192	183	1.0416
2012's eSur	187	167	8	64	179	0.9782
Embrace						
2009's eSur	175	80	1	79	174	1.02
2010's eSur	175	925	6	71	169	0.989
2011's eSur	175	5234	15	102	160	0.9477
2012's eSur	175	5234	15	111	160	0.9507
ObjectPut						
2009's eSur	non-participation					
2010's eSur						
2011's eSur	621	50	8	41	613	1.0006
2012's eSur	621	50	8	34	613	0.9983
CellToEar						
2009's eSur	non-participation					
2010's eSur						
2011's eSur						
2012's eSur	194	491	15	248	179	1.004
PersonRuns						
2009's eSur	107	356	5	351	102	1.068
2010's eSur	non-participation					
2011's eSur						
2012's eSur	107	785	9	181	98	0.9752
Pointing						
2009's eSur	non-participation					
2010's eSur						
2011's eSur	1063	2113	21	123	1042	1.0206
2012's eSur	1063	39120	32	74	1031	0.9942

5. Conclusion

This year we improved our system significantly in human detection where introduce geometric constraints, tracking where a robust tracking-by-detection approach with an optimized observation model is employed, and event detection where uneven-sequence classifier is used for the action-like event detection. The promising results of our system this year verify the effectiveness of these improvements. However, we believe there are still large improvement spaces for our system in exploring more effective and descriptive event models.

Reference

[1] Zhipeng Hu, Yaowei Wang, Yonghong Tian, Tiejun Huang, Selective Eigenbackgrounds Method for Background Subtraction in Crowed Scenes. ICIP 2010

[2]M. Andriluka, S. Roth, B. Schiele. People-tracking-by-detection and people-detection-by-tracking. Conference on Computer Vision and Pattern Recognition (CVPR), Page(s): 1–8, 2008.

[3] A. Garcia-Martin, A. Hauptmann, J.M. Martinez: People detection based on appearance and motion models. Advanced Video and Signal-Based Surveillance (AVSS), Page(s): 256 – 260, 2011

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[5]Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, ShaiAvidan: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. CVPR (2) 2006: 1491-1498

[6] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, Luc Van Gool. Online Multi-Person Tracking-by-Detection from a Single, Uncalibrated Camera. PAMI, 2010.

[7]Yasemin Altun, Loannis Tsochantaridis and Thomas Hofmann. Hidden Markov Support Vector Machines. ICML, 2003.

[8] J. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Statist. 29(5), 2001, 1189-1232.

[9] D. Reid, An algorithm for tracking multiple targets, IEEE Transactions on Automatic Control, Volume: 24, Issue: 6, 843 – 854, 1979

[10] I.J. Cox, S.L. Hingorani, An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 18, Issue: 2, 138 – 150, 1996

[11] H. Shimodaira, et al, Dynamic Time-Alignment Kernel in Support Vector Machine, Proc. Advances in Neural Information Processing Systems, 14, vol.2, pp.921-928, 2001.

[12] D. Mitzel, P. Sudowe, and B. Leibe. Real-Time Multi-PersonTracking with Time-Constrained Detection. In BMVC, 2011.

[13] A. Ess, B. Leibe, L.V. Gool. Depth and Appearance for Mobile Scene Analysis. In ICCV, 2007.

[14] Schindler, and L. Van Gool. Coupled detection and trajectory estimation for multi-object tracking. In ICCV,2007.

[15] S. Avidan. Ensemble tracking. PAMI, 29(2):261–271, 2007.

[16] B. Babenko, M. Yang, S. Belongie. Visual Tracking with Online Multiple Instance Learning. In CVPR, 2009.

[17] Sun, X., Hauptmann, Alexander, Action recognition via local descriptors and holistic features, Computer Science Department, 2009.

[18] R. Hess and A. Fern. Discriminatively Trained Particle Filters for Complex Multi-Object Tracking. In CVPR, 2009.

[19] H. Kuhn. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2:83–87, 1955.

[20] Y. Li, and J. Shawe-Taylor, The SVM with uneven margins and Chinese document categorization. In Proc. of PACLIC17, 2003, pp.216-227.

[21] M.-Y. Chen and A. Hauptmann, MoSIFT: Recognizing huamn actions in surveillance videos. Technical Report, CMU-CS-09-161, Carnegie Mellon University, 2009.

[22] Shimodaira, H, Dynamic time-alignment kernel in support vector machine, Proc. Advances in Neural Information Processing Systems 2, 2001, 921-928.

[23] T. Duong, H. Bui, D. Phung, and S. Venkatesh, Activity recognition and abnormality detection with the switching hidden semimarkov model, in IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[24] N. Smith and M. Niranjan, Data-dependent Kernels in SVM classication of speech patterns, in ICSLP,vol.1, 2000, pp.297-300.

[25] C. Watkins, Dynamic Alignment Kernels, in Advances in Large Margin Classifiers (A. J. Smola and P. L. Bartlett and B. Schölkopf and D. Schuurmans,ed.), ch. 3, 2000, pp. 39-50, The MIT Press.

[26] Hospedales, T., Li, J., Gong, S., Xiang, T, Identifying rare and subtle behaviours: A weakly supervised joint topic model, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, no. 99.

[27] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S, Behavior recognition via sparse spatio-temporal features, VS-PETS, 2005. [28] Zhou, Y., Yan, S., Huang, T.S, Pair-activity classification by bi-trajectories analysis, CVPR, 2008.